CSCI 4144/6405: Data Mining and Data Warehousing

Data Warehousing & Data Mining
Using Microsoft SQL Server 2012

(A Condensed Tutorial)

Updated by:
Mayank Malhotra

Faculty of Computer Science
Dalhousie University

January 2014

1 Introduction

This tutorial is a condensed and simplified version of Microsoft SQL Server 2012. It provides a brief introduction on how to use the basic tools of DW and DM for producing results for an application. If you need more details for some sections of this document please read the complete Microsoft SQL Server tutorial (http://technet.microsoft.com/en-us/library/ms170208.aspx).

2 Workstation Login

- Currently only workstations in Teaching Lab 1 and 2 have SQL Server 2012 installed. You can also borrow installation disks from Helpdesk and install the software on your personal computer. Microsoft SQL Server 2012 is also available for downloading on the following webpage: https://msdnaa.cs.dal.ca/. Please login to a workstation in Lab1/2 using your own CS account and password.

3 Open Business Intelligence Development Studio

Launch **Business Intelligence Development Studio** (**Start** menu → All Programs --> Microsoft SQL Server 2012 -->SQL server data tools-> SQL Server Business Intelligence Development Studio). If you are using the program for the first time, select Business Intelligence Settings in the pop-up window. In window 8 operating system you can also search for SQL server data tools, Please see **Figure 1**.

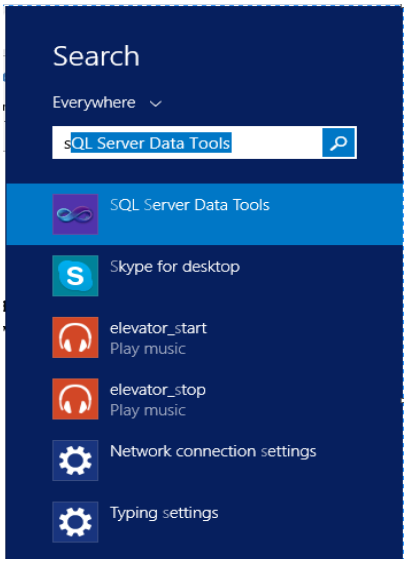

Figure 1: Open Business Intelligence Development Studio

4 Create/Open an Analysis Service Project

Creating a new project by using the Analysis Services Project template creates an empty project in which you define Analysis Services objects that you can then deploy to into a new Analysis Services database (or overwrite an existing Analysis Services database).

To create a new project with Business Intelligence Development Studio, by clicking File→New→Project, it will prompt you to the dialog to choose the **Analysis Server Project** template. In the **Location** item, choose the folder where you can store the project files for future access (Figure 2).

If you want to open an existing project, In Business Intelligence Development Studio, File→Open→Project, this will prompt you to the dialog to choose a project you have created before.

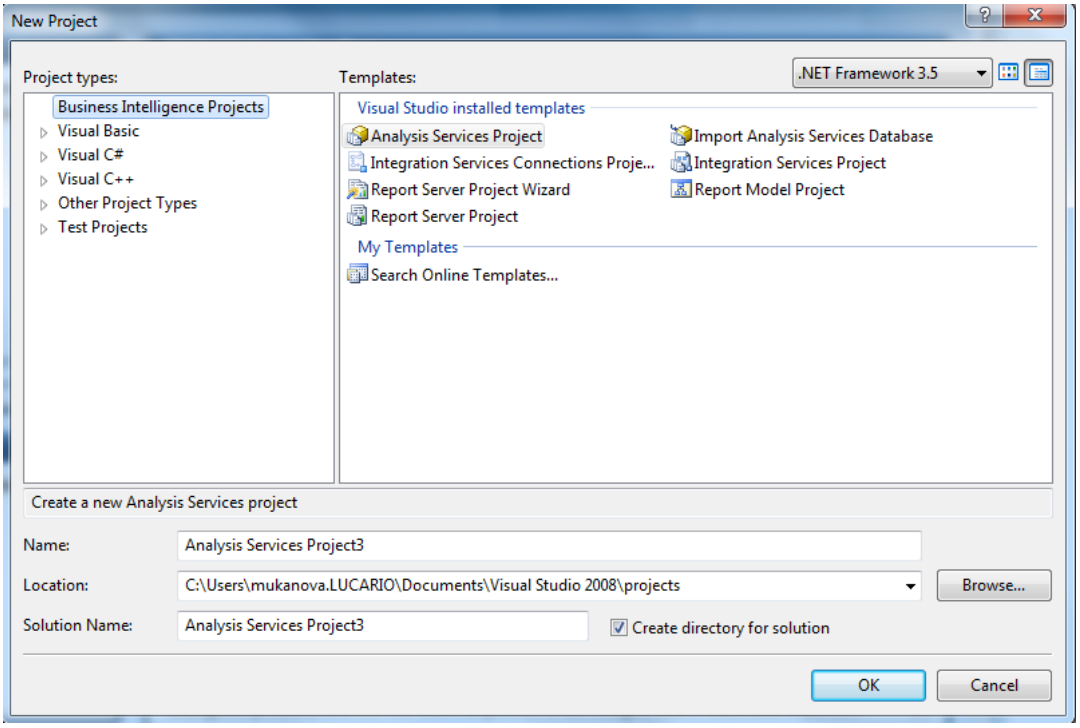

Figure 2: New analysis services project

5 Create Data Source

Data source needs to be specified in the created Analysis Service Project so that the analysis operations can be linked to the data source. Right click Data Sources in Solution Explorer, open **Data Source Wizard (**Figure 3**).**

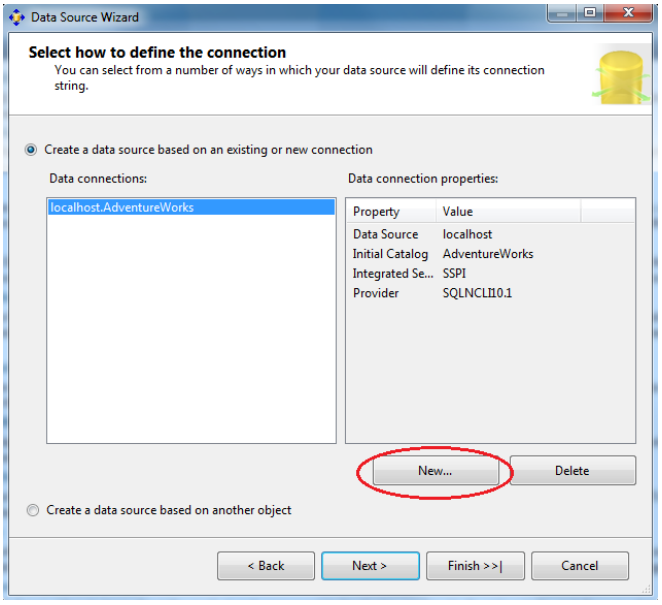

Figure 3: Data source wizard

Select "**New**" in the *Select how to define the connection*. This will open the **Connection Manager window** (Figure 4), where you should input the following information:

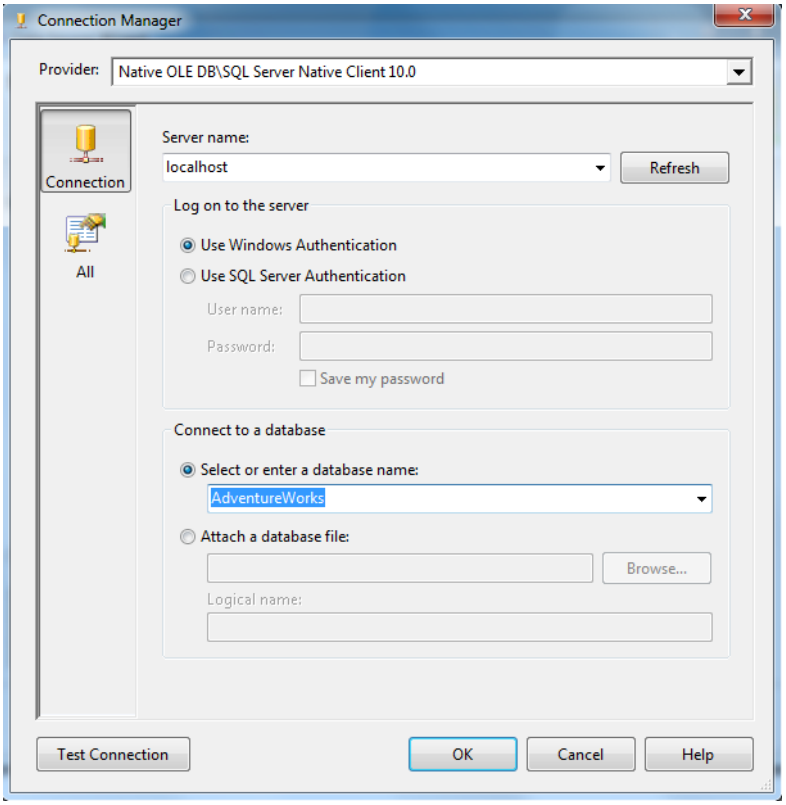

Figure 4: Connection Manager

Server name: select from the dropdown list or type in localhost (if database is installed on local machine), Click "**Test Connection**" to check after the server is picked.
Log on the server: choose "Use window authentication";
Connect to a database: choose "select or enter a database name" and from the list choose the example database "AdventureWorks" provided by Microsoft SQL Server 2012.

In the following **Impersonation Information** dialog, choose "use the service account".

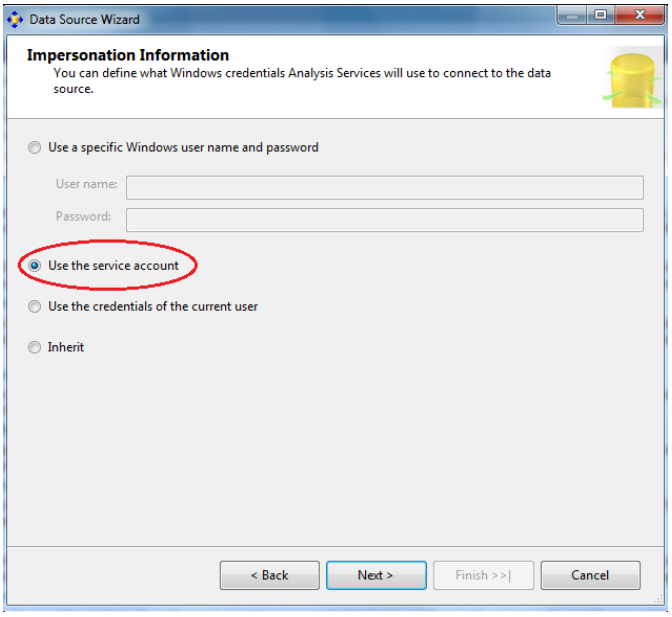

Figure 5: Impersonation Information

After finishing the Data Source Wizard, you can see AdventureWorks.ds appears under Data Sources folder in Solution Explorer.

NOTE: If you open an existing project on a different workstation, you may have to create a new data source.

6 Create Data Sources Viewer

You can see all (or part of) tables in the database and their relationship by creating Data Sources Viewer.

Right click Data Sources Views in Solution Explorer, open **Data Sources Views Wizard**.

In **Selection Tables and Views**, you can select the tables to be viewed. You also can use >> to select all the tables in the current database for viewing.

Note: Even though the assignment does not require you to create data source viewer, I strongly recommend you to do so before OLAP and data mining since it provides a convenient reference to relationships between tables in the database that may be useful in data warehousing and data mining.

7 Create Data Cube Objects

7.1 Cube Wizard

For OLAP analysis, you need to create data cubes. Right click Data Cube folder and click "new cube" in order to start the **Cube Wizard**.

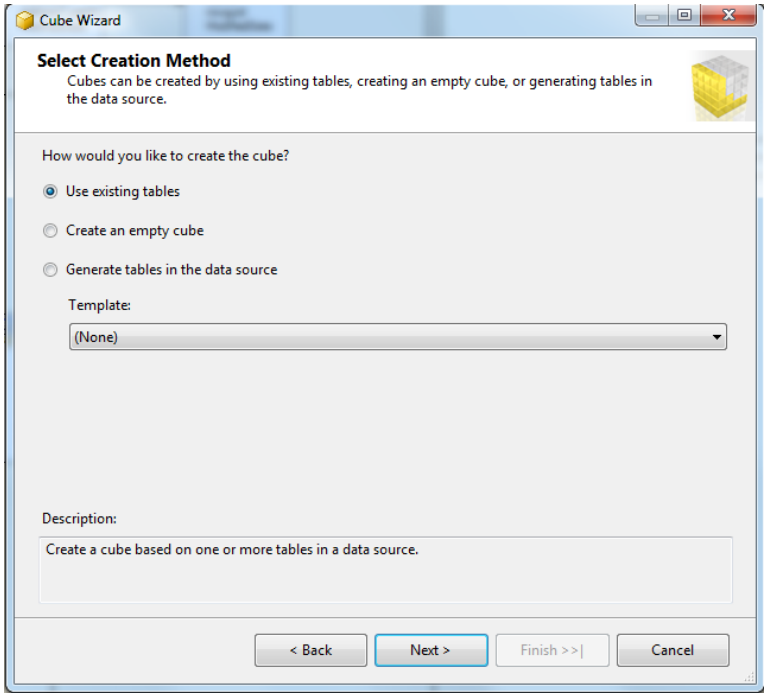

Figure 6: Cube Wizard

In **Select Creation Method**, choose "***Use existing tables***",

In **Select measure group tables** dialog**,** choose "AdventureWorks" as the Data source view;

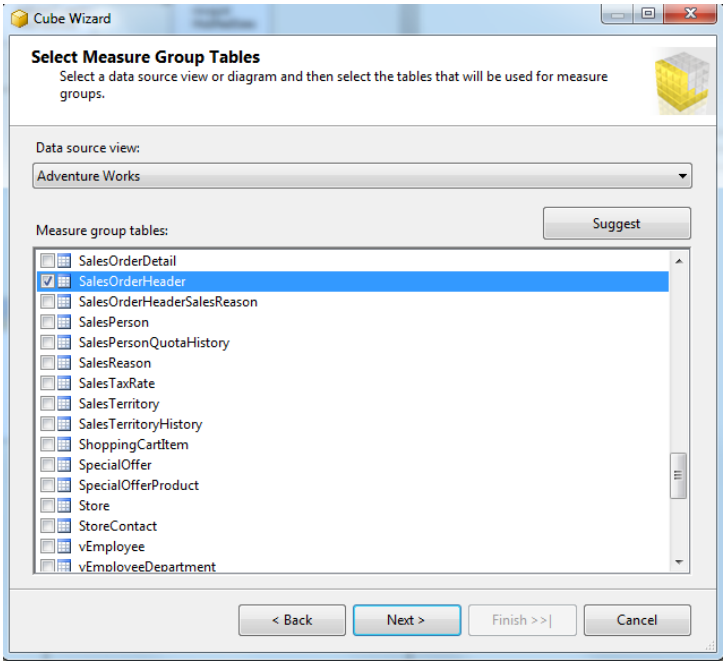

Figure 7: Select measure group tables in Cube Wizard

In **Measure group tables**, tick table "SalesOrderHeader" that will be used as the fact table.

In **Select Measures**, choose the measures you want to include in the cube.

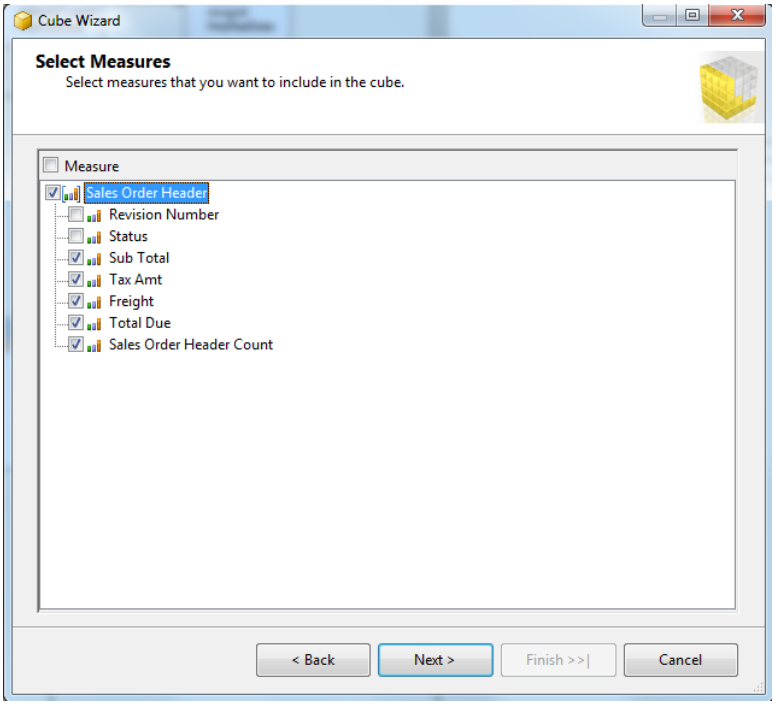

Figure 8: Select measures in Cube Wizard

In **Select New Dimensions,** tick the dimensions to be created, based on available tables.

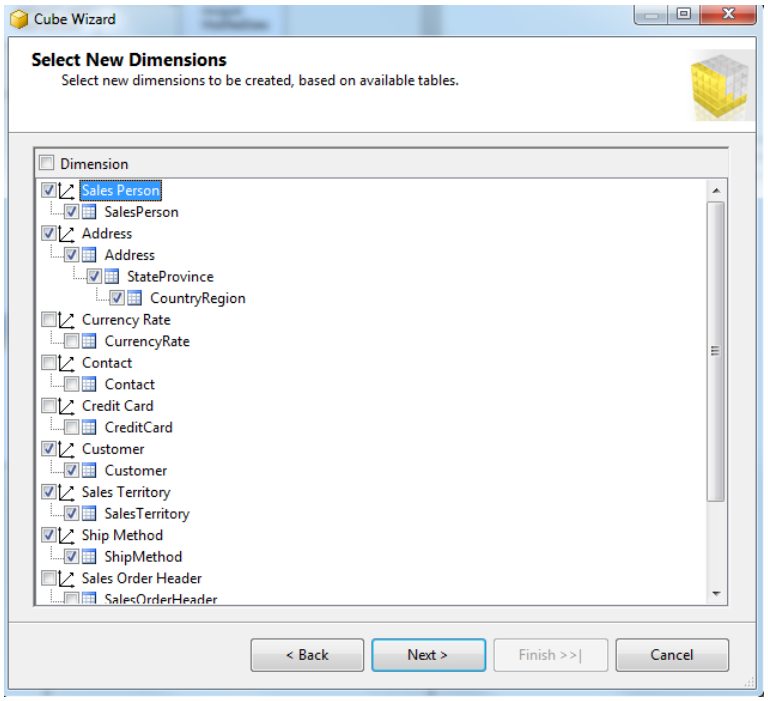

Figure 9: Select new dimensions Cube Wizard

Enter a Cube name to finish the cube wizard

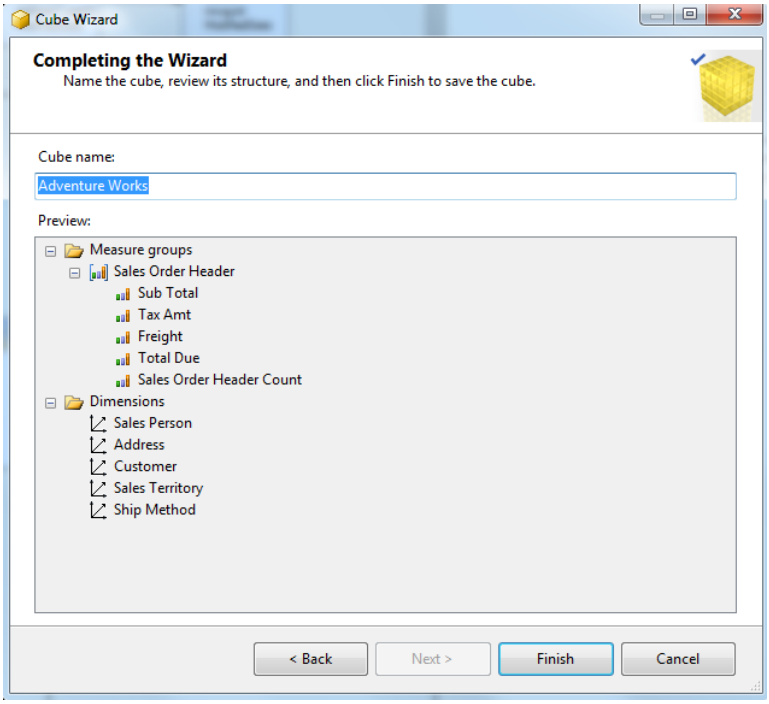

Figure 10: Finishing the Cube Wizard

7.2Dimension modification

After defining the initial cube, you are ready to improve the usefulness and friendliness of the cube. Double-click the dimension name in the Dimensions node of Solution Explorer, switch to Dimension Designer (Figure 11).

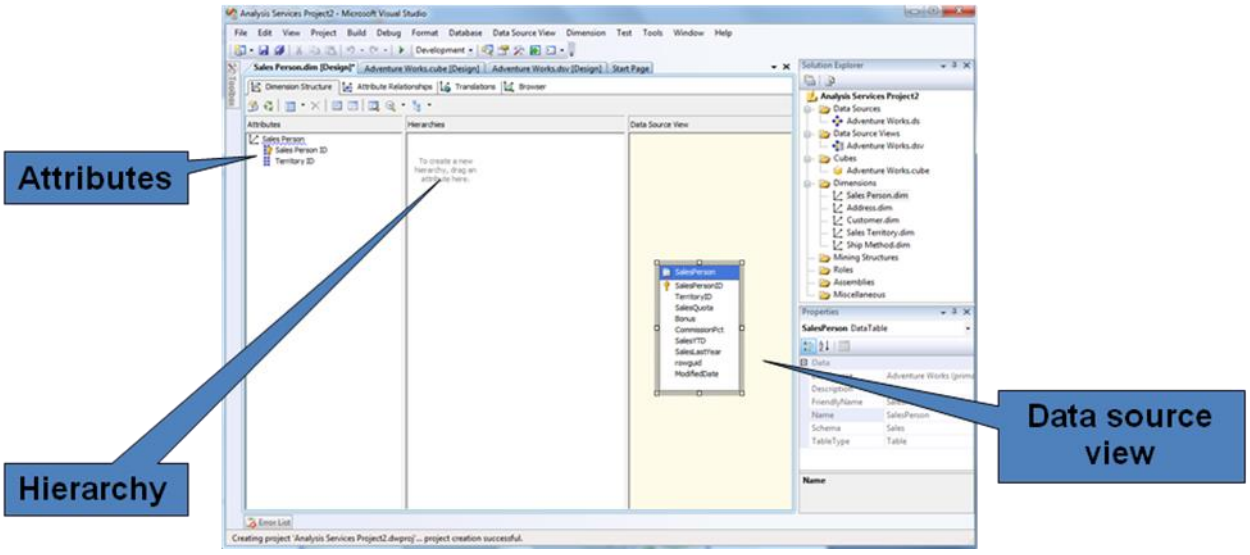

Figure 11: Dimension Designer

- Rename an attribute

  In the Attributes pane, right-click dimension name and select Rename.

- Add Attributes to a dimension

  Drag the table columns from the table in the Data Source View pane to the Attributes pane.

- Create a Hierarchy on a dimension

  Create a new hierarchy by dragging an attribute from the Attributes pane to the Hierarchies pane.

7.3 OLAP operators for DW

Before viewing the cube, you need to process both the Cubes object and Dimensions object. Right-click the Cube you have created and choose "**Process**" and start to "**run**" the processing of the cube object. Similarly, right-click Dimentions choose "**Process**" and start to "**run**" the processing of the dimension object.

After the processing of both cube and dimension objects, right-click the name of your cube object, say

it is called “Adventure Works”, and select “**Browse**”.

In left-most tree-like list, you can select the measures from the fact table for viewing. Right-click the measures you want to view and select “Add to data area”. The selected measures will be shown in the middle panel.

In the middle panel, you can specify the expression(s) for dimension attributes in order to view the measures that satisfy the given expression(s). For example, you can give the expression like “Shipmethod Equal 5” to view, for instance, the Total Due that using the 5th kind of ship method. In this way, you can easily perform the basic OLAP operations such as drill down, roll up, slice and dice for data exploration from the cube (Figure 12).

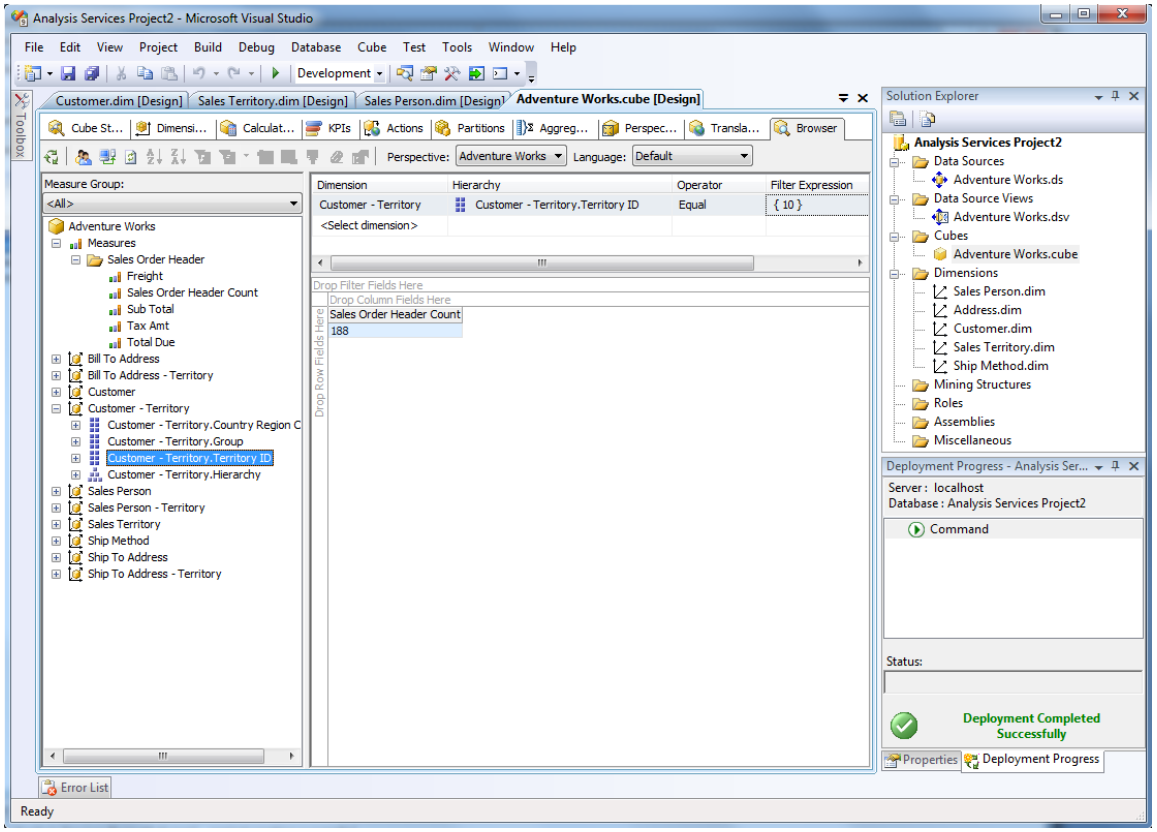

Figure 12: Cube data exploration and OLAP operations

8 Create Data Mining Objects

There are a number of data mining algorithms available in SQL Server 2012 such as Decision Trees, Association Rules, Clustering, Time Series and Text Mining, etc. When you build data mining models in Microsoft SQL Server 2012 Analysis Services (SSAS), the first step is to create a mining structure,

by using the Data Mining Wizard in Business Intelligence Development Studio. The mining structure defines the data domain from which mining models are built.

You can use **Data Mining Wizard** to create in one step the data mining objects. Right-click the **Mining Structure Folder** in the Solution Explorer and then choose New Mining Structure. On the first page, you choose whether you are creating a model from a relational or multidimensional source (an OLAP cube). In our case, select **"From existing relational database or data warehouse".**

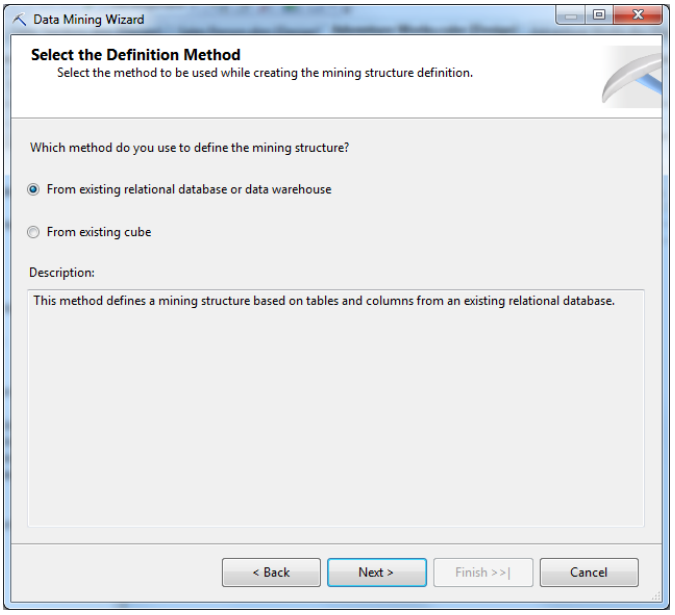

Figure 13: Select the definition method of the mining structure

The next page asks you which algorithm to use to create your initial mining model. Choosing which algorithm you are going to use is dependant on the business problem you are trying to solve. In our case, we're going to select **Microsoft Clustering** or **Decision Trees** algorithm for clustering and classification (**Figure 14**). In what follows, we will focus on these two data mining algorithms.

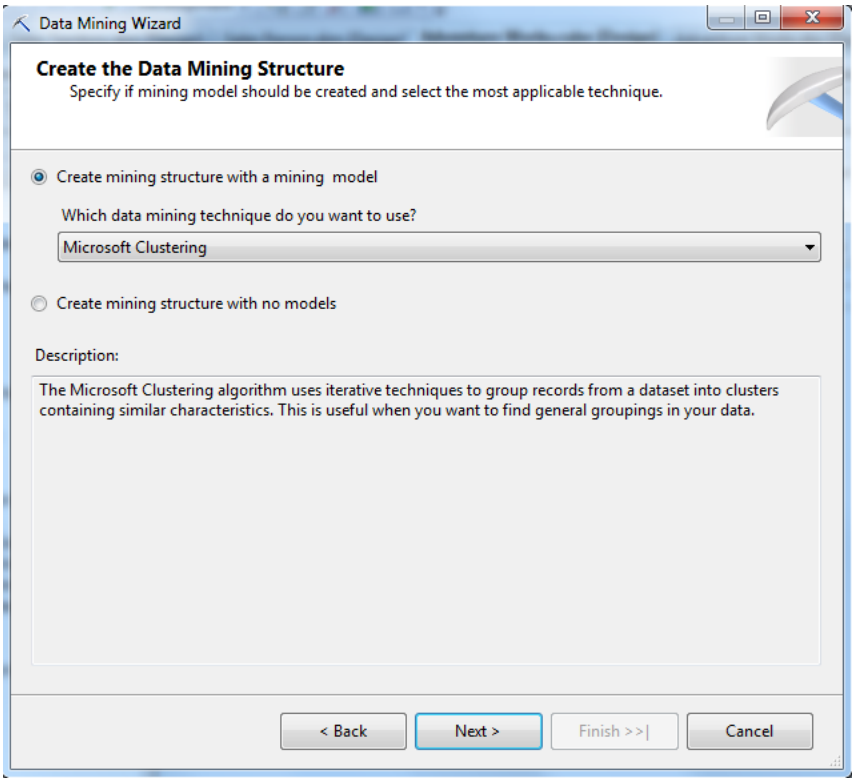

Figure 14: Choose a data mining algorithm to be applied

8.1.1 Microsoft Clustering

The next page after we choose "Microsoft Clustering",

- Select Data Source: Adventure Works

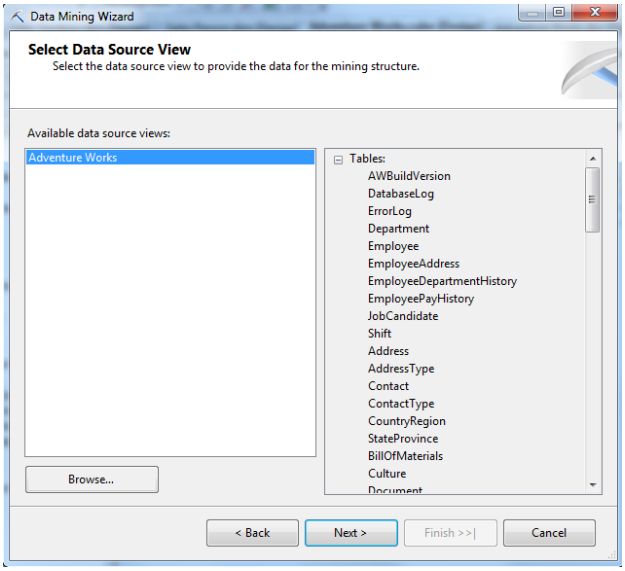

Figure 15: Select a data source for clustering

- Specify the table type
  - Select the case table from some defined views,
  - Whether the table is nested or not.
  - For example, select **SalesTerritory** table for clustering analysis.

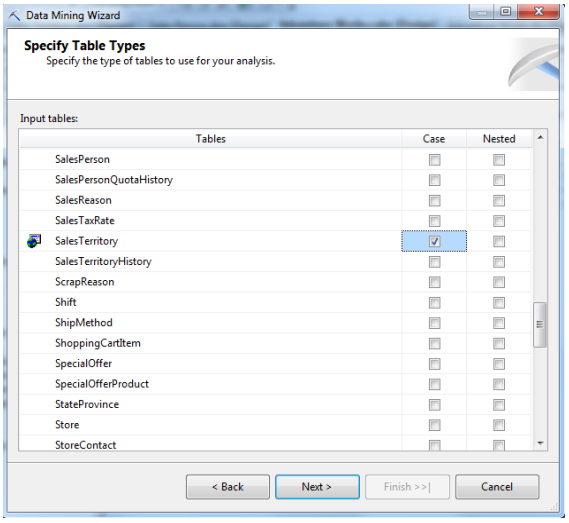

Figure 16: Select a case table for clustering

- Specify the Training data
  - Specify the key, input or predictable one from the attributes.
    - if we would like to **cluster territories based on their sales last year**,
      - **TerritoryID** is the key
      - **SalesLastYear** is the input for providing information to clustering.
    - Note that in the case of the clustering algorithm, there is no need for a predictable attribute.

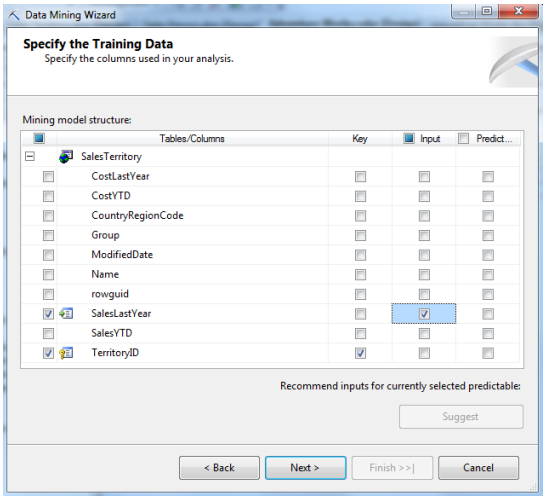

Figure 17: Specify the training data

- specify the column content and data type

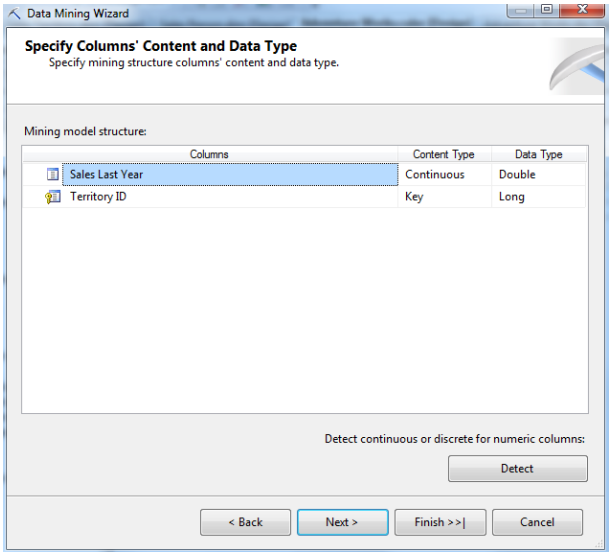

Figure 18: specify the column content and data type

- Create testing data set

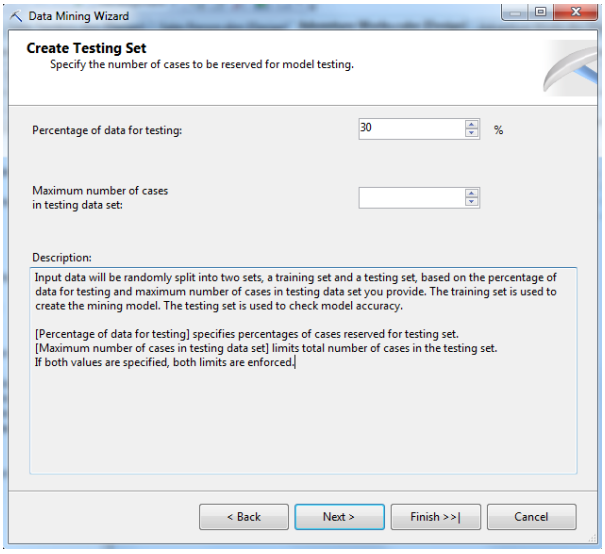

Figure 19: create testing data set

- Finally, set a descriptive name for the data mining structure and data mining model container.

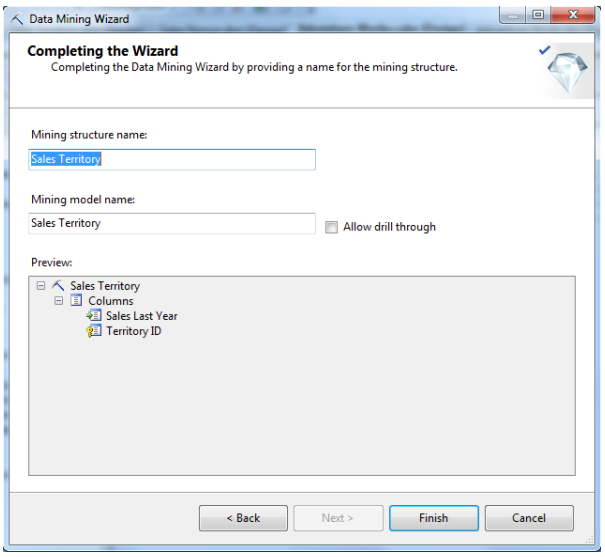

Figure 20: Name the structure and the model

8.1.2 The cluster viewer (Figure 21)

To process the clustering object, right click "**Process**". After clustering object is processed, we can view the clustering result through the cluster viewer. Right-click clustering object and select "**Browse**".

The Microsoft Cluster Viewer provides the following tabs for use in exploring clustering mining models:

- Cluster Diagram
- Cluster Profiles
- Cluster Characteristics
- Cluster Discrimination

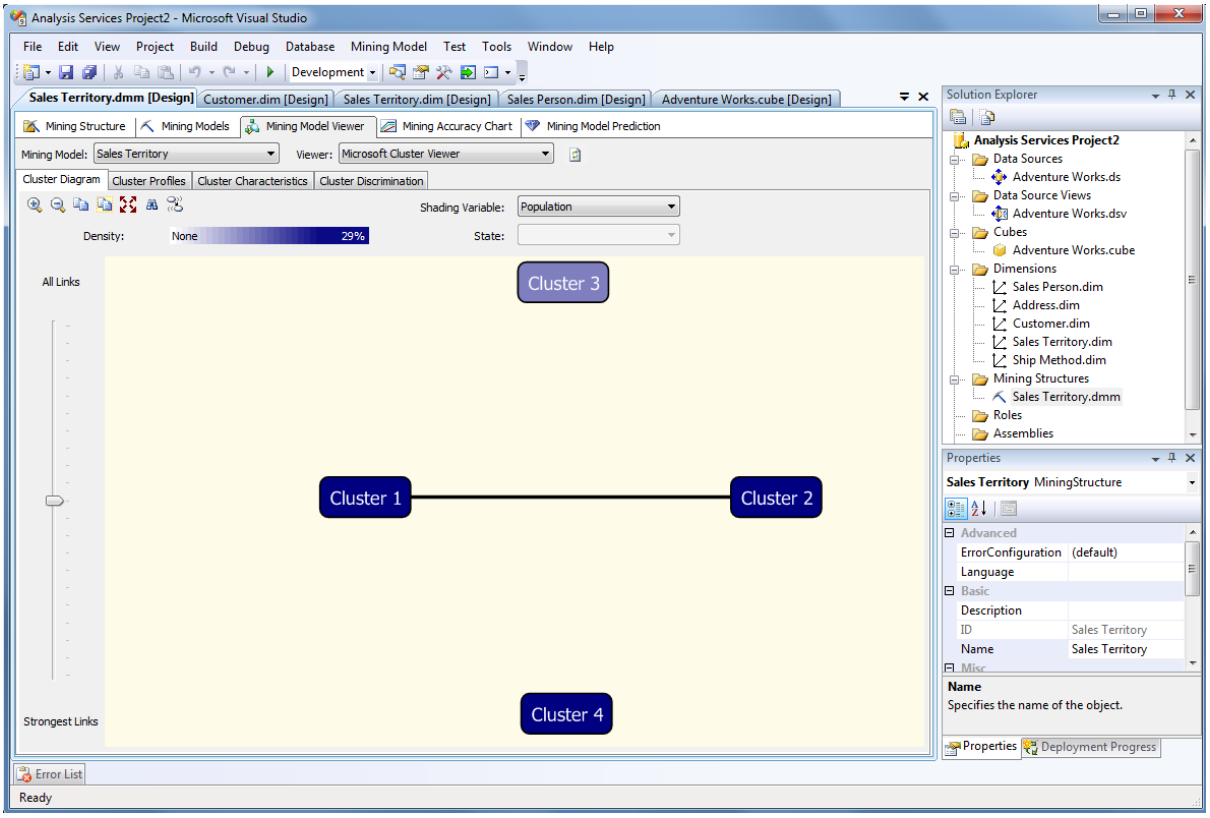

Figure 21: Browsing clustering results

Cluster Diagram

The Cluster Diagram tab of the Microsoft Cluster Viewer displays all the clusters that are in a mining model. The shading of the line that connects one cluster to another represents the strength of the similarity of the clusters. If the shading is light or nonexistent, the clusters are not very similar. As the line becomes darker, the similarity of the links becomes stronger. You can adjust how many lines the viewer shows by adjusting the slider to the right of the clusters. Lowering the slider shows only the strongest links.

By default, the shade represents the population of the cluster. By using the Shading Variable and State options, you can select which attribute and state pair the shading represents. The darker the shading, the greater the attribute distribution is for a specific state. The distribution decreases as the shading gets lighter.

To rename a cluster, right click its node and select Rename Cluster. The new name is persisted to the server.

To copy the visible section of the diagram to the Clipboard, right click the empty area of the Cluster Diagram, in the popup menu, click Copy Graph View. To copy the complete diagram, click Copy Entire Graph. You can also zoom in and out by using Zoom In and Zoom Out, or you can fit the diagram to the screen by using Scale Diagram to Fit in Window.

Cluster Profiles

The Cluster Profiles tab provides an overall view of the clusters that the algorithm in your model creates. This view displays each attribute, together with the distribution of the attribute in each cluster. An InfoTip for each cell displays the distribution statistics, and an InfoTip for each column heading displays the cluster population. Discrete attributes are shown as colored bars, and continuous attributes are shown as a diamond chart that represents the mean and standard deviation in each cluster. The Histogram bars option controls the number of bars that are visible in the histogram for discrete attributes. If more bars exist than you choose to display, the bars of highest importance are retained, and the remaining bars are grouped together into a gray bucket.

You can change the default names of the clusters, to make the names more descriptive. Rename a cluster by right-clicking its column heading and selecting Rename cluster. You can also hide clusters by selecting Hide column.

Cluster Characteristics

To use the Cluster Characteristics tab, select a cluster from the Cluster list. After you select a cluster, you can examine the characteristics that make up that specific cluster. The attributes that the cluster contains are listed in the Variables columns, and the state of the listed attribute is listed in the Values column. Attribute states are listed in order of importance, described by the probability that they will appear in the cluster. The probability is shown in the Probability column.

Cluster Discrimination

You can use the Cluster Discrimination tab to compare attributes between two clusters. Use the Cluster 1 and Cluster 2 lists to select the clusters to compare. The viewer determines the most important differences between the clusters, and displays the attribute states that are associated with the differences, in order of importance. A bar to the right of the attribute shows which cluster the state favors, and the size of the bar shows how strongly the state favors the cluster.

8.2 Microsoft Decision Tree

The steps to create decision tree object are pretty similar to those in creating clustering object outlined above. In the case of classification, we select SalesOrderHeader as the table for analysis. We would like to predict the total sale using two attributes, OnlineOrderFlag and TerritoryID by building a decision tree.

In this case, SaleOrderID attribute is the key for identifying each case, the OnlineOrderFlag and TerritoryID are used as the input attributes and TotalDue is used as the predictable attribute. **Note that in the case of the classification algorithm, you need to specify a predictable attribute.**

And finally the wizard ends after setting a descriptive name for the data mining structure and data mining model container.

Also, after the classification objects have been created, we need to process these objects to generate decision tree results. Right click the relevant data mining objects and select "**Process**".

8.2.1 The Tree Viewer (Figure 22)

The Microsoft Tree Viewer for viewing decision tree built includes the following tabs and panes:

- Decision Tree
- Dependency Network
- Mining Legend

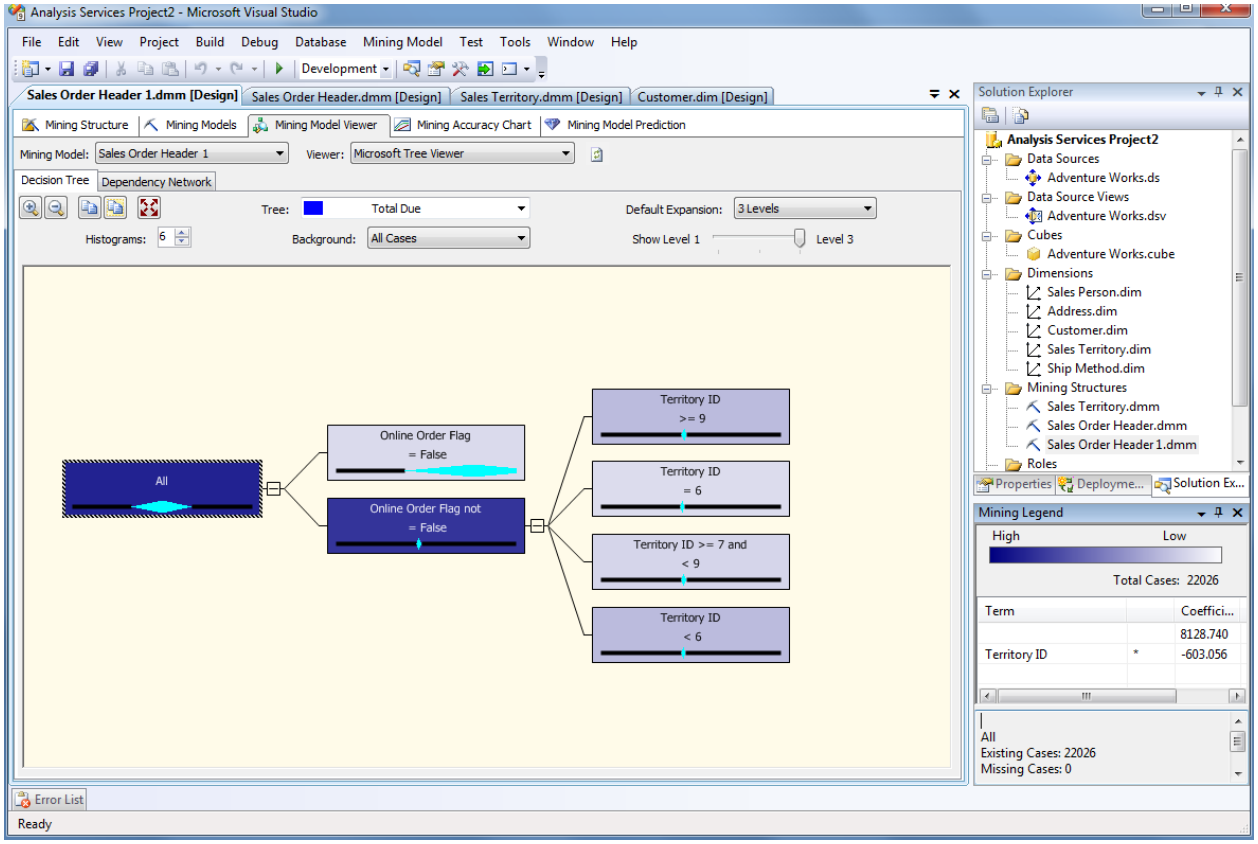

Figure 22: Results of Decision tree

Decision Tree

When you build a decision tree model, Analysis Services builds a separate tree for each predictable attribute. You can view an individual tree by selecting it from the **Tree** list on the **Decision Tree** tab of the viewer.

A decision tree is composed of a series of splits, with the most important split, as determined by the algorithm, at the left of the viewer in the **All** node. Additional splits occur to the right. The split in the **All** node is most important because it contains the strongest split-causing conditional in the dataset, and therefore it caused the first split.

You can expand or collapse individual nodes in the tree to show or hide the splits that occur after each node. You can also use the options on the **Decision Tree** tab to affect how the tree is displayed. Use the **Show Level** slider to adjust the number of levels that are shown in the tree. Use **Default Expansion** to set the default number of levels that are displayed for all trees in the model.

- ## Predicting Discrete Attributes

When a tree is built with a discrete predictable attribute, the viewer displays the following on each node in the tree:

- The condition that caused the split.
- A histogram that represents the distribution of the states of the predictable attribute, ordered by popularity.

You can use the **Histogram** option to change the number of states that appear in the histograms in the tree. This is useful if the predictable attribute has many states. The states appear in a histogram in order of popularity from left to right; if the number of states that you choose to display is fewer than the total number of states in the attribute, the least popular states are displayed collectively in gray. To see the exact count for each state for a node, pause the pointer over the node to view an InfoTip, or select the node to view its details in the **Mining Legend**.

The background color of each node represents the concentration of cases of the particular attribute state that you select by using the **Background** option. You can use this option to highlight nodes that contain a particular target in which you are interested.

- ## Predicting Continuous Attributes

When a tree is built with a continuous predictable attribute, the viewer displays a diamond chart, instead of a histogram, for each node in the tree. The diamond chart has a line that represents the range of the attribute. The diamond is located at the mean for the node, and the width of the diamond represents the variance of the attribute at that node. A thinner diamond indicates that the node can create a more accurate prediction. The viewer also displays the regression equation, which is used to determine the split in the node.

- ## Additional Decision Tree Display Options

You can use the zoom options on the **Decision Tree** tab to zoom in or out of a tree, or use **Size to Fit** to fit the whole model in the viewer screen. If a tree is too large to be sized to fit the screen, you can use the **Navigation** option to navigate through the tree. Clicking **Navigation** opens a separate navigation window that you can use to select sections of the model to display.

You can also copy the tree view image to the Clipboard, so that you can paste it into documents or into image manipulation software. Use **Copy Graph View** to copy only the section of the tree that is visible in the viewer, or use **Copy Entire Graph** to copy all the expanded nodes in the tree.

Dependency Network

The **Dependency Network** displays the dependencies between the input attributes and the predictable attributes in the model. The slider at the left of the viewer acts as a filter that is tied to the strengths of the dependencies. If you lower the slider, only the strongest links are shown in the viewer.

When you select a node, the viewer highlights the dependencies that are specific to the node. For example, if you choose a predictable node, the viewer also highlights each node that helps predict the predictable node.

Mining Legend

The **Mining Legend** displays the following information when you select a node in the decision tree model:

- The number of cases in the node, broken down by the states of the predictable attribute.
- The probability of each case of the predictable attribute for the node.
- A histogram that includes a count for each state of the predictable attribute.
- The conditions that are required to reach a specific node, also known as the *node path*.